6.231 DYNAMIC PROGRAMMING

LECTURE 14

LECTURE OUTLINE

- We start a ten-lecture sequence on advanced infinite horizon DP and approximation methods
- We allow infinite state space, so the stochastic shortest path framework cannot be used any more
- Results are rigorous assuming a finite or countable disturbance space
  - This includes deterministic problems with arbitrary state space, and countable state Markov chains
  - Otherwise the mathematics of measure theory make analysis difficult, although the final results are essentially the same as for finite disturbance space
- We use Vol. II of the textbook, starting with discounted problems (Ch. 1)
- The central mathematical structure is that the DP mapping is a contraction mapping (instead of existence of a termination state)

DISCOUNTED PROBLEMS/BOUNDED COST

- Stationary system with arbitrary state space

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots$$

- Cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$

$$J_\pi(x_0) = \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}$$

with $\alpha < 1$, and for some M, we have

$$|g(x, u, w)| \leq M, \qquad \forall\ (x, u, w)$$

- We have

$$\big|J_\pi(x_0)\big| \leq M + \alpha M + \alpha^2 M + \cdots = \frac{M}{1 - \alpha}, \quad \forall\ x_0$$

- The "tail" of the cost $J_\pi(x_0)$ diminishes to 0
- The limit defining $J_\pi(x_0)$ exists

WE ADOPT "SHORTHAND" NOTATION

- Compact pointwise notation for functions:
  - If for two functions J and J' we have $J(x) = J'(x)$ for all x, we write $J = J'$
  - If for two functions J and J' we have $J(x) \le J'(x)$ for all x, we write $J \le J'$
  - For a sequence $\{J_k\}$ with $J_k(x) \to J(x)$ for all x, we write $J_k \to J$; also $J^* = \min_\pi J_\pi$

- Shorthand notation for DP mappings (operate on functions of state to produce other functions)

$$(TJ)(x) = \min_{u \in U(x)} \mathop{E}_{w} \Big\{ g(x, u, w) + \alpha J\big(f(x, u, w)\big) \Big\}, \ \forall\, x$$

TJ is the optimal cost function for the one-stage problem with stage cost g and terminal cost αJ.

- For any stationary policy μ

$$(T_\mu J)(x) = \mathop{E}_{w} \Big\{ g\big(x, \mu(x), w\big) + \alpha J\big(f(x, \mu(x), w)\big) \Big\}, \ \forall\, x$$

- For finite-state problems:

$$T_\mu J = g_\mu + \alpha P_\mu J, \qquad TJ = \min_\mu T_\mu J$$

"SHORTHAND" COMPOSITION NOTATION

- Composition notation: T^2J is defined by $(T^2J)(x) = (T(TJ))(x)$ for all x (similar for T^kJ)
- For any policy $\pi = \{\mu_0, \mu_1, \ldots\}$ and function J:
  - $T_{\mu_0}J$ is the cost function of π for the one-stage problem with terminal cost function αJ
  - $T_{\mu_0}T_{\mu_1}J$ (i.e., T_{μ_0} applied to $T_{\mu_1}J$) is the cost function of π for the two-stage problem with terminal cost $\alpha^2 J$
  - $T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}J$ is the cost function of π for the N-stage problem with terminal cost $\alpha^N J$
- For any function J:
  - TJ is the optimal cost function of the one-stage problem with terminal cost function αJ
  - T^2J (i.e., T applied to TJ) is the optimal cost function of the two-stage problem with terminal cost $\alpha^2 J$
  - T^NJ is the optimal cost function of the N-stage problem with terminal cost $\alpha^N J$

"SHORTHAND" THEORY – A SUMMARY

- Cost function expressions [with $J_0(x) \equiv 0$]

$$J_\pi(x) = \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J_0)(x), \quad J_\mu(x) = \lim_{k\to\infty} (T_\mu^k J_0)(x)$$

- Bellman's equation: $J^* = TJ^*, \quad J_\mu = T_\mu J_\mu$
- Optimality condition:

$$\mu\text{: optimal} \quad <==> \quad T_\mu J^* = TJ^*$$

- Value iteration: For any (bounded) J and all x,

$$J^*(x) = \lim_{k\to\infty} (T^k J)(x)$$

- Policy iteration: Given μ^k:
  - Policy evaluation: Find J_{μ^k} by solving

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}$$

  - Policy improvement: Find μ^{k+1} such that

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}$$

SOME KEY PROPERTIES

- Monotonicity property: For any functions J and J' such that $J(x) \leq J'(x)$ for all x, and any μ

$$(TJ)(x) \leq (TJ')(x), \qquad \forall\, x,$$

$$(T_\mu J)(x) \leq (T_\mu J')(x), \qquad \forall\, x.$$

Also

$$J \leq TJ \quad \Rightarrow \quad T^k J \leq T^{k+1} J, \qquad \forall\, k$$

- Constant Shift property: For any J, any scalar r, and any μ

$$\big(T(J + re)\big)(x) = (TJ)(x) + \alpha r, \qquad \forall\, x,$$

$$\big(T_\mu(J + re)\big)(x) = (T_\mu J)(x) + \alpha r, \qquad \forall\, x,$$

where e is the unit function $[e(x) \equiv 1]$ (holds for most DP models).

- A third important property that holds for some (but not all) DP models is that T and T_μ are contraction mappings (more on this later).

CONVERGENCE OF VALUE ITERATION

- If $J_0 \equiv 0$,

$$J^*(x) = \lim_{N\to\infty} (T^N J_0)(x), \qquad \text{for all } x$$

Proof: For any initial state x_0, and policy $\pi = \{\mu_0, \mu_1, \ldots\}$,

$$\begin{aligned} J_\pi(x_0) &= E\left\{\sum_{k=0}^{\infty} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &= E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\quad + E\left\{\sum_{k=N}^{\infty} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \end{aligned}$$

from which

$$J_\pi(x_0) - \frac{\alpha^N M}{1-\alpha} \le (T_{\mu_0} \cdots T_{\mu_{N-1}} J_0)(x_0) \le J_\pi(x_0) + \frac{\alpha^N M}{1-\alpha},$$

where $M \ge |g(x, u, w)|$. Take the min over π of both sides. **Q.E.D.**

BELLMAN'S EQUATION

- The optimal cost function J^* satisfies Bellman's Eq., i.e. $J^* = TJ^*$.

Proof: For all x and N,

$$J^*(x) - \frac{\alpha^N M}{1-\alpha} \le (T^N J_0)(x) \le J^*(x) + \frac{\alpha^N M}{1-\alpha},$$

where $J_0(x) \equiv 0$ and $M \ge |g(x,u,w)|$.

- Apply T to this relation and use Monotonicity and Constant Shift,

$$(TJ^*)(x) - \frac{\alpha^{N+1} M}{1-\alpha} \le (T^{N+1} J_0)(x) \le (TJ^*)(x) + \frac{\alpha^{N+1} M}{1-\alpha}$$

- Take limit as $N \to \infty$ and use the fact

$$\lim_{N\to\infty} (T^{N+1} J_0)(x) = J^*(x)$$

to obtain $J^* = TJ^*$. **Q.E.D.**

THE CONTRACTION PROPERTY

- Contraction property: For any bounded functions J and J', and any μ,

$$\max_x \left|(TJ)(x) - (TJ')(x)\right| \leq \alpha \max_x \left|J(x) - J'(x)\right|,$$

$$\max_x \left|(T_\mu J)(x) - (T_\mu J')(x)\right| \leq \alpha \max_x \left|J(x) - J'(x)\right|.$$

Proof: Denote $c = \max_{x \in S} \left|J(x) - J'(x)\right|$. Then

$$J(x) - c \leq J'(x) \leq J(x) + c, \qquad \forall\ x$$

Apply T to both sides, and use the Monotonicity and Constant Shift properties:

$$(TJ)(x) - \alpha c \leq (TJ')(x) \leq (TJ)(x) + \alpha c, \qquad \forall\ x$$

Hence

$$\left|(TJ)(x) - (TJ')(x)\right| \leq \alpha c, \qquad \forall\ x.$$

Similar for T_μ. **Q.E.D.**

IMPLICATIONS OF CONTRACTION PROPERTY

- We can strengthen our earlier result:
- Bellman's equation $J = TJ$ has a unique solution, namely J^*, and for any bounded J, we have

$$\lim_{k\to\infty} (T^k J)(x) = J^*(x), \qquad \forall\, x$$

Proof: Use

$$\max_x \big|(T^k J)(x) - J^*(x)\big| = \max_x \big|(T^k J)(x) - (T^k J^*)(x)\big| \leq \alpha^k \max_x \big|J(x) - J^*(x)\big|$$

- Special Case: For each stationary μ, J_μ is the unique solution of $J = T_\mu J$ and

$$\lim_{k\to\infty} (T_\mu^k J)(x) = J_\mu(x), \qquad \forall\, x,$$

for any bounded J.

- Convergence rate: For all k,

$$\max_x \big|(T^k J)(x) - J^*(x)\big| \leq \alpha^k \max_x \big|J(x) - J^*(x)\big|$$

NEC. AND SUFFICIENT OPT. CONDITION

- A stationary policy μ is optimal if and only if $\mu(x)$ attains the minimum in Bellman's equation for each x; i.e.,

$$TJ^* = T_\mu J^*.$$

Proof: If $TJ^* = T_\mu J^*$, then using Bellman's equation ($J^* = TJ^*$), we have

$$J^* = T_\mu J^*,$$

so by uniqueness of the fixed point of T_μ, we obtain $J^* = J_\mu$; i.e., μ is optimal.

- Conversely, if the stationary policy μ is optimal, we have $J^* = J_\mu$, so

$$J^* = T_\mu J^*.$$

Combining this with Bellman's equation ($J^* = TJ^*$), we obtain $TJ^* = T_\mu J^*$. **Q.E.D.**

COMPUTATIONAL METHODS - AN OVERVIEW

- Typically must work with a finite-state system. Possibly an approximation of the original system.

- Value iteration and variants
  - Gauss-Seidel and asynchronous versions

- Policy iteration and variants
  - Combination with (possibly asynchronous) value iteration
  - "Optimistic" policy iteration

- Linear programming

$$\text{maximize} \quad \sum_{i=1}^{n} J(i)$$

$$\text{subject to} \quad J(i) \leq g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j), \quad \forall\ (i,u)$$

- Versions with subspace approximation: Use in place of $J(i)$ a low-dim. basis function representation, with state features $\phi_m(i)$, $m = 1, \ldots, s$

$$\tilde{J}(i,r) = \sum_{m=1}^{s} r_m \phi_m(i)$$

and modify the basic methods appropriately.

USING Q-FACTORS I

- Let the states be $i = 1, \ldots, n$. We can write Bellman's equation as

$$J^*(i) = \min_{u \in U(i)} Q^*(i, u) \qquad i = 1, \ldots, n,$$

where

$$Q^*(i, u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + \alpha J^*(j)\big)$$

for all (i, u)

- $Q^*(i, u)$ is called the optimal Q-factor of (i, u)
- Q-factors have optimal cost interpretation in an "augmented" problem whose states are i and (i, u), $u \in U(i)$ - the optimal cost vector is (J^*, Q^*)
- The Bellman Eq. is $J^* = TJ^*$, $Q^* = FQ^*$ where

$$(FQ^*)(i, u) = \sum_{j=1}^{n} p_{ij}(u) \left(g(i, u, j) + \alpha \min_{v \in U(j)} Q^*(j, v) \right)$$

- It has a unique solution.

USING Q-FACTORS II

- We can equivalently write the VI method as

$$J_{k+1}(i) = \min_{u \in U(i)} Q_{k+1}(i,u), \qquad i = 1,\ldots,n,$$

where Q_{k+1} is generated for all i and $u \in U(i)$ by

$$Q_{k+1}(i,u) = \sum_{j=1}^{n} p_{ij}(u) \left(g(i,u,j) + \alpha \min_{v \in U(j)} Q_k(j,v) \right)$$

or $J_{k+1} = TJ_k$, $Q_{k+1} = FQ_k$.

- Equal amount of computation ... just more storage.

- Having optimal Q-factors is convenient when implementing an optimal policy on-line by

$$\mu^*(i) = \min_{u \in U(i)} Q^*(i,u)$$

- Once $Q^*(i,u)$ are known, the model [g and $p_{ij}(u)$] is not needed. Model-free operation.

- Stochastic/sampling methods can be used to calculate (approximations of) $Q^*(i,u)$ [not $J^*(i)$] with a simulator of the system.